LAMP-TR-149 February 2008

Generalizing Word Lattice Translation

Christopher Dyer, Smaranda Muresan, Philip Resnik

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
redpony, smara, resnik AT umd.edu

Abstract

Word lattice decoding has proven useful in spoken language translation; we argue that it provides a compelling model for translation of text genres, as well. We extend lattice decoding to hierarchical phrase-based models, providing a unified treatment with phrase-based decoding by treating lattices as a case of weighted finite-state automata. In the process, we resolve a significant complication that lattice representations introduce in reordering models. Our experiments evaluating the approach demonstrate substantial gains for Chinese-English and Arabic-English translation.

Keywords: word lattice translation, phrase-based and hierarchical models, statistical machine translation

The support of this research by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **FEB 2008**	2. REPORT TYPE	3. DATES COVERED **00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE **Generalizing Word Lattice Translation**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Institute for Advanced Computer Studies,College Park,MD,20742-3275**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Word lattice decoding has proven useful in spoken language translation; we argue that it provides a compelling model for translation of text genres, as well. We extend lattice decoding to hierarchical phrase-based models, providing a unified treatment with phrase-based decoding by treating lattices as a case of weighted finite-state automata. In the process, we resolve a significant complication that lattice representations introduce in reordering models. Our experiments evaluating the approach demonstrate substantial gains for Chinese-English and Arabic-English translation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

Generalizing Word Lattice Translation

Christopher Dyer, Smaranda Muresan, Philip Resnik
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
redpony, smara, resnik AT umd.edu

Abstract

Word lattice decoding has proven useful in spoken language translation; we argue that it provides a compelling model for translation of text genres, as well. We extend lattice decoding to hierarchical phrase-based models, providing a unified treatment with phrase-based decoding by treating lattices as a case of weighted finite-state automata. In the process, we resolve a significant complication that lattice representations introduce in reordering models. Our experiments evaluating the approach demonstrate substantial gains for Chinese-English and Arabic-English translation.

1 Introduction

When Brown and colleagues introduced statistical machine translation in the early 1990s, their key insight — harkening back to Weaver in the late 1940s – was that translation could be viewed as an instance of noisy channel modeling (Brown et al., 1990). They introduced a now standard decomposition that distinguishes modeling sentences in the target language (language models) from modeling the relationship between source and target language (translation models). Today, virtually all statistical translation systems seek the best hypothesis e for a given input f in the source language, according to

$$\hat{e} = \arg\max_{e} Pr(e|f) \tag{1}$$

An exception is the translation of speech recognition output, where the acoustic signal generally underdetermines the choice of source word sequence f. There, Bertoldi and others have recently found that, rather than translating a single-best transcription f, it is advantageous to allow the MT decoder to consider all possibilities for f by encoding the alternatives compactly as a confusion network or lattice (Bertoldi et al., 2007; Bertoldi and Federico, 2005; Koehn et al., 2007).

Why, however, should this advantage be limited to translation from spoken input? Consider: even for text, there are often multiple ways to derive a sequence of words from the input string. Segmentation of Chinese, decompounding in German, morphological analysis for Arabic — across a wide range of source languages, ambiguity in the input gives rise to multiple possibilities for the source word sequence. Nonetheless, state-of-the-art systems commonly identify a single analysis f during a preprocessing step, and decode according to the decision rule in (1).

In this paper, we go beyond speech translation by showing that lattice decoding can also yield improvements for text by preserving alternative analyses of the input. In addition, we generalize lattice decoding algorithmically, extending it for the first time to hierarchical phrase-based translation (Chiang, 2005; Chiang, 2007).

Formally, the approach we take can be thought of as a "noisier channel", where an observed signal o gives rise to a set of source-language strings $f' \in \mathcal{F}(o)$ and we seek

$$\begin{aligned} \hat{e} &= \arg\max_{e} \max_{f' \in \mathcal{F}(o)} Pr(e, f'|o) & (2) \\ &= \arg\max_{e} \max_{f' \in \mathcal{F}(o)} Pr(e) Pr(f'|e, o) & (3) \\ &= \arg\max_{e} \max_{f' \in \mathcal{F}(o)} Pr(e) Pr(f'|e) Pr(o|f'). & (4) \end{aligned}$$

Following Och and Ney (2002), we use the maximum entropy framework (Berger et al., 1996) to directly model the posterior $Pr(e, f'|o)$ with parameters tuned to minimize a loss function representing the quality only of the resulting translations. Thus, we make use of the following general decision rule:

$$\hat{e} = \arg\max_{e} \max_{f' \in \mathcal{F}(o)} \sum_{m=1}^{M} \lambda_m \phi_m(e, f', o) \qquad (5)$$

In principle, one could decode according to (2) simply by enumerating and decoding each $f' \in \mathcal{F}(o)$; however, for any interestingly large $\mathcal{F}(o)$ this will be impractical. We assume that for many interesting cases of $\mathcal{F}(o)$, there will be identical substrings that express the same content, and therefore that a lattice representation is appropriate.

In Section 2, we discuss decoding with this model in general, and then show how two widely used classes of translation model can be easily adapted for a lattice translation framework; we achieve a unified treatment of finite-state and hierarchical phrase-based models by treating lattices as a subcase of weighted finite state automata (FSAs). In Section 3, we identify and solve issues that arise with reordering in non-linear FSAs, i.e. FSAs where every path does not pass through every node. Section 4 presents two applications of the noisier channel paradigm, demonstrating substantial performance gains in Arabic-English and Chinese-English translation. In Section 5 we discuss relevant prior work, and we conclude in Section 6.

2 Decoding

Most statistical machine translation systems model translational equivalence using either finite state transducers or synchronous context free grammars (Lopez, to appear 2008). In this section we discuss the issues associated with adapting decoders from both classes of formalism to process word lattices. The first decoder we present is a SCFG-based decoder similar to the one described in Chiang (2007). The second is a phrase-based decoder implementing the model of Koehn et al. (2003).

2.1 Word lattices

A word lattice $\mathcal{G} = \langle V, E \rangle$ is a directed acyclic graph that formally is a weighted finite state automaton (FSA). We further stipulate that exactly one node has no out-going edges and is designated the 'end node'. Figure 1 illustrates three classes of word lattices.

A word lattice is useful for our purposes because it permits any finite set of strings to be represented and allows for substrings common to multiple members of the set to be represented with

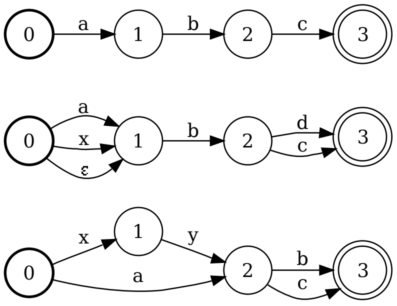

Figure 1: Three examples of word lattice: (a) sentence, (b) confusion network, and (c) general word lattice.

a single piece of structure. Additionally, all paths from one node to another form an equivalence class representing, in our model, alternative expressions of the same underlying communicative intent.

For translation, we will find it useful to encode $\mathcal{G}$ in a chart based on a topological ordering of the nodes, as described by Cheppalier et al. (1999). The nodes in the lattices shown in Figure 1 are labeled according to an appropriate numbering.

The chart-representation of the graph is a triple of 2-dimensional matrices $\langle \mathbf{F}, \mathbf{p}, \mathbf{R} \rangle$, which can be constructed from the numbered graph. $\mathbf{F}_{i,j}$ is the word label of the j^{th} transition leaving node i. The corresponding transition cost is $\mathbf{p}_{i,j}$. $\mathbf{R}_{i,j}$ is the node number of the node on the *right* side of the j^{th} transition leaving node i. Note that $\mathbf{R}_{i,j} > i$ for all i, j. Table 1 shows the word graph from Figure 1 represented in matrix form as $\langle \mathbf{F}, \mathbf{p}, \mathbf{R} \rangle$.

0			1			2		
a	1	1	b	1	2	c	1	3
a	$\frac{1}{3}$	1	b	1	2	c	$\frac{1}{2}$	3
x	$\frac{1}{3}$	1				d	$\frac{1}{2}$	3
ϵ	$\frac{1}{3}$	1						
x	$\frac{1}{2}$	1	y	1	2	b	$\frac{1}{2}$	3
a	$\frac{1}{2}$	**2**				c	$\frac{1}{2}$	3

Table 1: Topologically ordered chart encoding of the three lattices in Figure 1. Each cell ij in this table is a triple $\langle \mathbf{F}_{ij}, \mathbf{p}_{ij}, \mathbf{R}_{ij} \rangle$

2.2 Parsing word lattices

Chiang (2005) introduced hierarchical phrase-based translation models, which are formally based on synchronous context-free grammars (SCFGs). Translation proceeds by parsing the input using the source language side of the grammar, simultaneously building a tree on the target language side via the target side of the synchronized rules. Since decoding is equivalent to parsing, we begin by presenting a parser for word lattices, which is a generalization of a CKY parser for lattices given in Cheppalier et al. (1999).

Following Goodman (1999), we present our lattice parser as a deductive proof system in Figure 2. The parser consists of two kinds of items, the first with the form $[X \rightarrow \alpha \bullet \beta, i, j]$ repre-

Axioms:

$$\frac{}{[X \to \bullet\gamma, i, i] : w} \quad (X \xrightarrow{w} \langle\gamma,\ \alpha\rangle) \in G, i \in [0, |V|-2]$$

Inference rules:

$$\frac{[X \to \alpha \bullet \mathbf{F}_{j,k}\beta, i, j] : w}{[X \to \alpha\mathbf{F}_{j,k} \bullet \beta, i, \mathbf{R}_{j,k}] : w \times \mathbf{p}_{j,k}}$$

$$\frac{[X \to \alpha \bullet \beta, i, j] : w}{[X \to \alpha \bullet \beta, i, \mathbf{R}_{j,k}] : w \times \mathbf{p}_{j,k}} \quad \mathbf{F}_{j,k} = \epsilon$$

$$\frac{[Z \to \alpha \bullet X\beta, i, k] : w_1 \quad [X \to \gamma\bullet, k, j] : w_2}{[Z \to \alpha X \bullet \beta, i, j] : w_1 \times w_2}$$

Goal state:

$$[S \to \gamma\bullet, 0, |V|-1]$$

Figure 2: Word lattice parser for an unrestricted context free grammar G.

senting rules that have yet to be completed and span node i to node j. The other items have the form $[X, i, j]$ and indicate that non-terminal X spans $[i, j]$. As with sentence parsing, the goal is a deduction that covers the spans of the entire input lattice $[S, 0, |V|-1]$.

The three inference rules are: 1) match a terminal symbol and move across one edge in the lattice 2) move across an ϵ-edge without advancing the dot in an incomplete rule 3) advance the dot across a non-terminal symbol given appropriate antecedents.

Using memoization of previously encountered items, this parser runs in polynomial time.

2.3 From parsing to MT decoding

A target language model is necessary to generate fluent output. To do so, the grammar is intersected with an n-gram LM. To mitigate the effects of the combinatorial explosion of non-terminals this entails, a pruning strategy is necessary. We use *cube-pruning* (Chiang, 2007).

2.4 Lattice translation with FSTs

A second important class of translation models includes those based formally on FSTs. We present a description of the decoding process for a word lattice using a representative FST model, the phrase-based translation model described in Koehn et al. (2003).

Phrase-based models translate a foreign sentence f into the target language e by breaking up f into a sequence of phrases $\overline{f}_1^I$, where each phrase $\overline{f}_i$ can contain 1 or more contiguous words and is translated into a target phrase e_i of 1 or more contiguous words. Each word in f must be translated exactly once. To generalize this model to word lattices, it is necessary to choose both a path through the lattice and a partitioning of the sentence this induces into a sequence of phrases $\overline{f}_1^I$. Although the number of source phrases in a word lattice can be exponential in the number of nodes in the lattice, enumerating the *possible translations* of every span in a lattice is in practice tractable, as described by Bertoldi et al. (2007).

2.5 Decoding with phrase-based models

We adapted the Moses phrase-based decoder to translate word lattices (Koehn et al., 2007). The unmodified decoder builds a translation hypothesis from left to right by selecting a range of un-

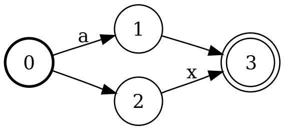

Figure 3: The span $[0, 3]$ has one inconsistent covering, $[0, 1] + [2, 3]$.

translated words and adding translations of this phrase to the end of the hypothesis being extended. When no untranslated words remain, the translation process is complete.

The word lattice decoder works similarly, only now the decoder keeps track of the number of nodes that have been covered, given a topological ordering of the nodes. For example, assuming the third lattice in Figure 1 is our input, if the edge with word *a* is translated, this will cover *two* untranslated nodes [0,1] in the coverage vector, even though it is only a single word. As with sentence-based decoding, a translation hypothesis is complete when all nodes in the input lattice are covered.

2.6 Non-monotonicity and unreachable nodes

The changes described thus far are straightforward adaptations of the underlying phrase-based sentence decoder; however, dealing properly with non-monotonic decoding of word lattices introduces some minor complexity that is worth mention. In the sentence decoder, any translation of any span of untranslated words is an allowable extension of a partial translation hypothesis, provided that the coverage vectors of the extension and the partial hypothesis do not intersect. In a non-linear word lattice, a further constraint must be enforced ensuring that there is always a path from the starting node of the translation extension's source to the node representing the nearest right edge of the already-translated material, as well as a path from the ending node of the translation extension's source to future translated spans. Figure 3 illustrates the problem. If [0,1] is translated, the decoder must not consider translating [2,3] as a possible extension of this hypothesis since there is no path from node 1 to node 2 and therefore the span [1,2] would never be covered. In the parser that forms the basis of the hierarchical decoder described in Section 2.3, no such restriction is necessary since grammar rules are processed in a strictly left-to-right fashion without any skips.

3 Distortion in a non-linear word lattice

In both hierarchical and phrase-based models, the distance between words in the source sentence is used to limit where in the target sequence their translations will be generated. In phrase based translation, distortion is modeled explicitly. Models that support non-monotonic decoding generally include a distortion penalty, such as $|a_i - b_{i-1} - 1|$ where a_i is the starting position of the foreign phrase $\overline{f}_i$ and b_{i-1} is the ending position of phrase $\overline{f}_{i-1}$ (Koehn et al., 2003). The intuition behind this model is that since most translation is monotonic, the cost of skipping ahead or back in the source should be proportional to the number of words that are skipped. Additionally, a maximum distortion limit is used to restrict the size of the search space.

In linear word graphs, such as confusion networks, the distance metric used for the distortion penalty and for distortion limits is well defined; however, in a non-linear word graph, it poses the problem illustrated in Figure 4. Assuming the left-to-right decoding strategy described in the

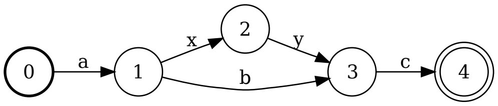

Figure 4: Distance-based distortion problem. How far is it from node 4 to node 0?

previous section, if *c* is generated by the first target word, the distortion penalty associated with "skipping ahead" should be either 3 or 2, depending on what path is chosen to translate the span [0,3]. In large lattices, this problem can be quite significant. The cost of swapping two adjacent words in one path can grow arbitrarily large and may ultimately be impossible because of the distortion limit in certain lattice structures.

Although hierarchical phrase-based models do not model distortion explicitly, Chiang (2007) suggests using a span length limit to restrict the window in which reordering can take place.[1] The decoder enforces the constraint that a synchronous rule learned from the training data (the only mechanism by which reordering can be introduced) can span maximally Λ words in f. As for distortion limits, this limit is also poorly defined for non-linear lattices.

Since we want a distance metric that will restrict as few local reorderings as possible on *any* path, we use a function $\xi(a, b)$ returning the length of the shortest path between nodes a and b. Since this function is not dependent on the exact path chosen, it can be computed in advance of decoding using an all-pairs shortest path algorithm (Cormen et al., 1989).

3.1 Experimental results

We tested the effect of the distance metric on translation quality using Chinese word segmentation lattices (Section 4.1, below) using both a hierarchical and phrase-based system modified to translate word lattices. We compared the shortest-path distance metric with a baseline which uses the difference in node number as the distortion distance. For an additional datapoint, we added a lexicalized reordering model that models the probability of each phrase pair appearing in three different orientations (swap, monotone, other) in the training corpus (Koehn et al., 2005).

Table 2 summarizes the results of the phrase-based systems. On both test sets, the shortest path metric improved the BLEU scores. As expected, the lexicalized reordering model improved translation quality over the baseline; however, the improvement was more substantial in the model that used the shortest-path distance metric (which was an already higher baseline). Table 3 summarizes the results of our experiment comparing the performance of two distance metrics to determine whether a rule has exceeded the decoder's span limit. The pattern is the same, showing a clear increase in BLEU for the shortest path metric over the baseline.

4 Exploiting Source Language Alternatives

Chinese word segmentation. A necessary first step in translating Chinese using standard models is segmenting the character stream into a sequence of words. Word-lattice translation offers two possible improvements over the conventional approach. First, a lattice may represent multiple

[1] This is done to reduce the size of the search space and because hierarchical phrase-based translation models are inaccurate models of long-distance distortion.

Distance metric	MT05	MT06
Difference	29.43	27.86
Difference+LexRO	29.74	28.90
ShortestP	29.93	28.65
ShortestP+LexRO	30.72	29.92

Table 2: Effect of distance metric on phrase-based model performance.

Distance metric	MT05	MT06
Difference	30.63	29.57
ShortestP	31.76	30.43

Table 3: Effect of distance metric on hierarchical model performance.

alternative segmentations of a sentence; input represented in this way will be more robust to errors made by the segmenter. [2] Second, features from the target side of the training corpus can be used to build an optimal segmentation of the training data. For example, a hypothesized Chinese word consisting of two characters that also has a high probability of having a fertility of 2 might be a good candidate for splitting into two separate words.[3] Figure 5 illustrates a lattice based on three different segmentations.

Arabic morphological variation. Arabic orthography is problematic for lexical and phrase-based MT approaches since a large class of functional elements (prepositions, pronouns, tense markers, conjunctions, definiteness markers) are attached to their host stems. Thus, while the training data may provide good evidence for the translation of a particular stem by itself, the same stem may not be attested when attached to a particular conjunction. The general solution taken is to take the best possible morphological analysis of the text (it is often ambiguous whether a piece of a word is part of the stem or merely a neighboring functional element), and then make a subset of the bound functional elements in the language into freestanding tokens. Figure 6 illustrates the surface form of Arabic orthography as well as a morphological segmentation. A possible problem with this approach is that as the amount and variety of training data increases, the optimal segmentation strategy changes: more aggressive segmentation results in fewer OOV tokens, but automatic evaluation metrics indicate lower translation quality, presumably because the smaller units are being translated less idiomatically (Habash and Sadat, 2006). Lattices allow the decoder to attempt to use the idiomatic surface forms but back off gracefully to a more aggressively segmented form of the text when there is insufficient evidence for a good translation of the surface tokens. Furthermore, since morphological analysis is an inherently ambiguous process, word lattices can effectively capture the resulting ambiguity.

4.1 Chinese Word Segmentation Experiments

In our experiments we used two state-of-the-art Chinese word segmenters: one developed at Harbin Institute of Technology (Zhao et al., 2001), and one developed at Stanford University (Tseng et al., 2005). In addition, we used a character-based segmentation. In the remaining of this paper, we use

[2] The segmentation process is ambiguous, even for native speakers of Chinese

[3] This is reminiscent of the approach taken by Ma et al. (2007), but permits a hypothesized Chinese word to be segmented differently in different contexts.

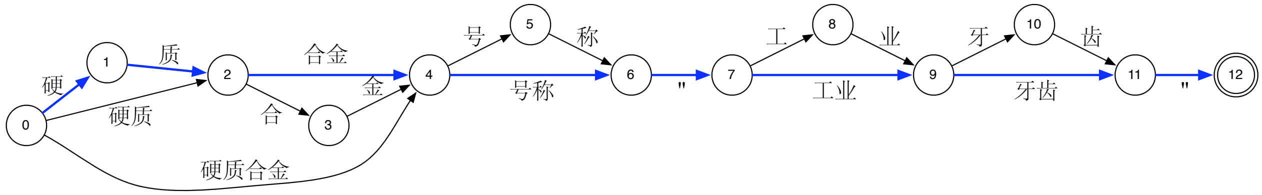

Figure 5: Sample Chinese segmentation lattice using three segmentations.

surface	wxlAl ftrp AlSyf kAn mEZm AlDjyj AlAElAmy m&ydA llEmAd .
segmented	w- xlAl ftrp Al- Syf kAn mEZm Al- Djyj Al- AElAmy m&ydA l- Al- EmAd .
(English)	During the summer period , most media buzz was supportive of the general .

Figure 6: Example of Arabic morphological segmentation.

cs for character segmentation, hs for Harbin segmentation and ss for Stanford segmentation. We built two types of lattices: one that combines the Harbin and Stanford segmenters (hs+ss), and one which uses all three segmentations (hs+ss+cs).

Data and Settings. The systems used in these experiments were trained on the NIST MT06 Eval corpus without the UN data (approximatively 950K sentences). The corpus has been segmented with the three segmentations. For the systems using word-lattices, we added the training data corresponding to each segmentation used in the source lattice. A trigram English language model with modified Kneser-Ney smoothing (Kneser and Ney, 1995) was trained on the English side of our training data as well as portions of the GigaWord v2 English Corpus, and was used for all experiments. The NIST MT03 test set was used as development set for optimizing the interpolation weights using minimum error rate training (Och, 2003). The testing was done on NIST 2005 and 2006 evaluation sets (MT05, MT06).

Experimental results: Word-lattices improve translation quality. We used both a phrase-based translation model, Moses (Koehn et al., 2007) and a hierarchical phrase-based translation model, Hiero (Chiang, 2005; Chiang, 2007). These two translation models illustrate global theoretical contributions presented in Section 2 and Section 3.

Since we are using a relatively small parallel training corpus and no additional lexical resources, the coverage of named entities (NEs) is rather poor. This is particularly problematic when using character-based segmentation. In order not to unfairly penalize the lattice system using the character segmentation, we did not character-segment NEs when generating the segmentation lattices. For this, we used a Chinese NE tagger (Florian et al., 2004), and only excluded NEs of type PERSON.

The results are presented in Table 4. We test statistical significance using bootstrap resampling (Koehn, 2004; Zhang et al., 2004).

Using word lattices improves BLEU scores both in the phrase-based model and hierarchical model as compared to the single best segmentation. All results using our word-lattice decoding for the hierarchical models (hs+ss and hs+ss+cs) are significantly better than the best segmentation (ss) ($p<0.05$ and $p<0.01$, respectively). For the phrase-based model, we obtain significant gains using our word-lattice decoder using all three segmentations on MT05 ($p<0.01$). The other results, while better than the best segmentation (hs) by at least 0.3 BLEU points, are not statistically significant. Even if the results are not statistically significant for MT06, there is a high decrease in

Moses (Source Type)	MT05 BLEU	MT06 BLEU
cs	0.2833	0.2694
hs	0.2905	0.2835
ss	0.2894	0.2801
hs+ss	0.2938	0.2870
hs+ss+cs	0.2993	0.2865
hs+ss+cs.lexRo	0.3072	0.2992

(a) Phrase-based Model

Hiero (Source Type)	MT05 BLEU	MT06 BLEU
cs	0.2904	0.2821
hs	0.3008	0.2907
ss	0.3071	0.2964
hs+ss	0.3132	0.3006
hs+ss+cs	0.3176	0.3043

(b) Hierachical Model

Table 4: Chinese Word Segmentation Results

Moses (Source Type)	MT05 BLEU	MT06 BLEU
surface	46.82	35.12
morh	50.87	38.41
morph+surface	52.25	40.08

(a) Phrase-based Model

Hiero (Source Type)	MT05 BLEU	MT06 BLEU
surface	52.53	39.91
morph	53.77	41.80
morph+sourface	54.53	42.87

(b) Hierachical model

Table 5: Arabic Morphology Results

OOV items when using word-lattices. For example, for MT06 the number of OOVs in the hs translation is 484. The number of OOVs decreased by 19% for hs+ss and by 75% for hs+ss+cs. As mentioned in Section 3, using lexical reordering for word-lattices further improves the translation quality, a statistically significant gain (p<0.01) for both MT05 and MT06 (Table 4(a)).

4.2 Arabic Morphology Experiments

In our experiments we used a fairly aggressive segmentation and normalization of Arabic text, most similar to the EN scheme described by Habash and Sadat (2006).

Data and Settings. For these experiments we subsampled the NIST MT08 training data (including UN).[4] For the morphology system, the size of the resulting subsampled data was 42M tokens for Arabic and 37M tokens for English. For the surface system, the size of the subsampled data was 25M tokens for Arabic and 30M tokens for English. To generate the translation model used in the word-lattice system, we extracted rules from a corpus with two versions of the source. For all systems, we used a 5-gram English LM trained on the non-UN part of the entire training data, plus the portions of the UN data present in the morph/surface subsamples, and 45M words from English GigaWord Corpus v3. The NIST MT03 test set was used as development set for optimizing the interpolation weights using minimum error rate training (Och, 2003). Testing was done on NIST 2005 and 2006 evaluation sets (MT05, MT06).

Experimental results: Word-lattices improve translation quality. Results are presented in Table 5. Using word-lattices to combine the surface forms with morphologically segmented forms improves BLEU scores both in the phrase-based and hierarchical models. All improvements are statistically significant ($p < .01$).

[4] We used a subsampling method proposed by Kishore Papineni, personal communication, that aims to include training sentences containing n-grams in the test data.

5 Prior work

Lattice Translation. The 'noisier channel' model of machine translation has been widely used in spoken language translation as an alternative to selecting the 1-best hypothesis from an ASR system and translating it (Ney, 1999; Casacuberta et al., 2004; Zhang et al., 2005; Saleem et al., 2005; Matusov et al., 2005; Bertoldi et al., 2007; Mathias, 2007). Several authors (e.g. Saleem et al. (2005) and Bertoldi et al. (2007)) comment directly on the impracticality of using n-best lists to translate speech.

Although translation is fundamentally a non-monotonic relationship between most language pairs, reordering has tended to be a secondary concern to the researchers who have worked on lattice translation. Matusov et al. (2005) decodes monotonically and then uses a finite state reordering model on the 1-best translation, along the lines of Bangalore and Riccardi (2000). Mathias (2007) and Saleem et al. (2004) only report results of monotonic decoding of the systems they describe. Bertoldi et al. (2007) solve the problem by requiring that their input be in the format of a confusion network, which enables the standard distortion penalty to be used. Finally, the system described by Zhang et al. (2005) use IBM Model 4 features to translate lattices. For the distortion model, they use an approach similar to ours where they use the maximum probability value over all possible paths in the lattice for each jump considered.

Applications of source lattices outside of the domain of spoken language translation have been far more limited. Costa-jussà and Fonollosa (2007) take steps in this direction by using lattices to encode multiple reorderings of the source language. Dyer (2007) uses confusion networks to encode morphological alternatives in Czech-English translation.

The Arabic-English morphological segmentation lattices are similar in spirit to backoff translation models (Yang and Kirchhoff, 2006) which consider alternative morphological segmentations and simplifications when a surface form is not found.

Parsing and formal language theory. There has been considerable work on parsing word lattices, much of it for language modeling applications in speech recognition (Ney, 1991; Cheppalier and Rajman, 1998; Hall and Johnson, 2003). Additionally, Grune and Jacobs (2008) describes an algorithm for intersecting an arbitrary FSA (of which word lattices are a subset) with a CFG. Klein and Manning (2001) formalizes parsing as a hypergraph search problem and presents an $O(n^3)$ CFG parser for FSAs.

6 Conclusions

We have achieved substantial gains in translation performance by decoding compact representations of alternative source language analyses, rather than single-best representations. Our results generalize previous gains for lattice translation of spoken language input, and we have further generalized the approach by introducing an algorithm for lattice decoding using a hierarchical phrase-based model. Furthermore we have shown that although word lattices complicate modeling of word reordering, a simple heuristic offers good performance and enables standard distortion models to be used with lattice input.

At this stage, we suggest taking an even more radical step. Up to this point, we have assumed that the ambiguity encoded in lattices is primarily the fault of imperfect analyzers; if a perfect analyzer could give us true single-best paths, we would simply use them in training and decoding. However, there is a deeper observation to be made here, namely that *the source language sentence* *is not the only way that the author's meaning could have been expressed*. The "noisier channel," implemented using lattice decoding, makes it natural to conceive of the source sentence as just one possible realization of the speaker's intended meaning. We conjecture that *introducing* ambiguity, by expanding that single realization into a lattice of possible alternative realizations, will improve the translation of what was *said* by bringing the system's input representations closer to what was *meant*.

Acknowledgments

This research was supported by the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-2-0001. The authors wish to thank Niyu Ge for the Chinese named-entity analysis, Pi-Chuan Chang and the Stanford NLP group for their assistance with the Stanford Chinese segmenter, and Tie-Jun Zhao and Congui Zhu for making the Harbin Chinese segmenter available to us.

References

S. Bangalore and G. Riccardi. 2000. Finite state models for lexical reordering in spoken language translation. In *Proc. Int. Conf. on Spoken Language Processing*, pages 422–425, Beijing, China.

A.L. Berger, V.J. Della Pietra, and S.A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71.

N. Bertoldi and M. Federico. 2005. A new decoder for spoken language translation based on confusion networks. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, Cancun, Mexico, December.

N. Bertoldi, R. Zens, and M. Federico. 2007. Speech translation by confusion network decoding. In *Proceeding of ICASSP 2007*, Honolulu, Hawaii, April.

N. Bertoldi, R. Zens, M. Federico, and W. Shen. 2007 (submitted). Efficient speech translation through confusion network decoding.

P.F. Brown, J. Cocke, S. Della-Pietra, V.J. Della-Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16:79–85, June.

F. Casacuberta, H. Ney, F. J. Och, E. Vidal, J. M. Vilar, S. Barrachina, I. Garcia-Varea, D. Llorens, C. Martinez, S. Molau, F. Nevado, M. Pastor, D. Pico, A. Sanchis, and C. Tillmann. 2004. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech & Language*, 18(1):25–47, January.

J. Cheppalier and M. Rajman. 1998. A generalized CYK algorithm for parsing stochastic CFG. In *Proceedings of the Workshop on Tabulation in Parsing and Deduction (TAPD98)*, pages 133–137, Paris, France.

J. Cheppalier, M. Rajman, R. Aragues, and A. Rozenknop. 1999. Lattice parsing for speech recognition. In *Sixth Conference sur le Traitement Automatique du Langage Naturel (TANL'99)*, pages 95–104.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, Michigan, June. Association for Computational Linguistics.

D. Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

T.H. Cormen, C. E. Leiserson, and R. L. Rivest, 1989. *Introduction to Algorithms*, pages 558–565. The MIT Press and McGraw-Hill Book Company.

M. Costa-jussà and J.A.R. Fonollosa. 2007. Analysis of statistical and morphological classes to generate weighted reordering hypotheses on a statistical machine translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 171–176, Prague.

C. Dyer. 2007. Noisier channel translation: translation from morphologically complex languages. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, June.

R. Florian, H. Hassan, A. Ittycheriah, H. Jing, N. Kambhatla, X. Luo, N Nicolov, and S Roukos. 2004. A statistical model for multilingual entity detection and tracking. In *Proc. of HLT-NAACL 2004*, pages 1–8.

J. Goodman. 1999. Semiring parsing. *Computational Linguistics*, 25:573–605.

D. Grune and C.J. H. Jacobs. 2008. Parsing as intersection. *Parsing Techniques*, pages 425–442.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proc. of NAACL*, New York.

K. Hall and M. Johnson. 2003. Language modeling using efficient best-first bottom-up parsing. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 507–512.

D. Klein and C. D. Manning. 2001. Parsing with hypergraphs. In *Proceedings of IWPT 2001*.

R. Kneser and H. Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of IEEE Internation Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL 2003*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 iwslt speech translation evaluation. In *Proceedings of IWSLT 2005*, Pittsburgh.

P. Koehn, H. Hoang, A. Birch Mayne, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177–180, Jun.

P. Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*.

A. Lopez. to appear 2008. Statistical machine translation. *ACM Computing Surveys*.

L. Mathias. 2007. *Statistical Machine Translation and Automatic Speech Recognition under Uncertainty*. Ph.D. thesis, The Johns Hopkins University.

E. Matusov, S. Kanthak, and H. Ney. 2005. On the integration of speech recognition and statistical machine translation. In *Proceedings of Interspeech 2005*.

H. Ney. 1991. Dynamic programming parsing for context-free grammars in continuous speech recognition. *IEEE Transactions on Signal Processing*, 39(2).

H. Ney. 1999. Speech translation: Coupling of recognition and translation. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 517–520, Phoenix, AR, March.

F. Och and H. Ney. 2002. Discriminitive training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 295–302.

S. Saleem, S.-C. Jou, S. Vogel, and T. Schulz. 2005. Using word lattice information for a tighter coupling in speech translation systems. In *Proceedings of ICSLP*, Jeju Island, Korea, Oct.

H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. 2005. A conditional random field word segmenter. In *Fourth SIGHAN Workshop on Chinese Language Processing*.

M. Yang and K. Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of the EACL 2006*, pages 41–48.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of LREC 2004*.

R. Zhang, G. Kikui, H. Yamamoto, and W. Lo. 2005. A decoding algorithm for word lattice translation in speech translation. In *Proceedings of the 2005 International Workshop on Spoken Language Translation*.

T. Zhao, L. Yajuan, Y. Muyun, and Y. Hao. 2001. Increasing accuracy of chinese segmentation with strategy of multi-step processing. In *Journal of Chinese Information Processing (Chinese Version)*, volume 1, pages 13–18.